Modeling Early Vocal Development Through Infant–Caregiver Interaction: A Review

Minoru Asada, *Fellow, IEEE*

***Abstract*—The developmental origin of language communication seems to involve vocal interactions between an infant and a caregiver, and one of the big mysteries related to this is how an infant learns to vocalize the caregiver's native language. Many theories attempt to explain this ability of infant as imitation based on acoustic matching. However, the acoustic qualities of speech produced by the infant and caregiver are quite different and therefore cannot be fully explained by imitation. Instead, the interaction itself may have an important role to play, but the mechanism is still unclear. In this paper, we review studies addressing this topic based on explicit interaction mechanisms using computer simulations and/or real vocal robots. The relationships between these approaches are analyzed after a brief review of the early development of an infant's speech perception and articulation based on observational studies in developmental psychology and a few neuroscientific imaging studies. Finally, future issues related to real infant–caregiver vocal interaction are outlined.**

***Index Terms*—Vocal development, social interaction, affirmative bias, virtual and physical agents.**

I. Introduction

LANGUAGE is a unique communication capability of the human species because it provides a powerful means of referencing for objects, events, and relationships. From an evolutionary perspective, it is still a great mystery as to how human beings acquired language. Moreover, the way in which human infants and children learn to use language has yet to be fully understood from a developmental perspective [1]. In this paper, we focus on the developmental aspect of language communication. The origin of language communication seems to be the vocal interactions between an infant and its caregiver, with an important question being how the infant learns to vocalize the native language of the caregiver.

Computational modeling has been applied to explain the developmental process of speech perception and articulation. Some of this modeling has not explicitly addressed the issue of interaction between an infant and its caregiver (see [2]–[10]). Others have attempted to explain an infant's imitation ability based on acoustic matching (see [11] and [12]). However, the acoustic features of the vocalizations produced by an infant and its caregiver are quite different and therefore cannot adequately explain the acoustic imitation. Rather, the interactions themselves appear to play an important role. To investigate how infant–caregiver interactions affect early vocal development, we review modeling approaches [13]–[18] based on explicit interaction mechanisms using computer simulations and/or real vocal robots. These are called constructive approaches, and cognitive developmental robotics [19], [20] have been advocating the need to identify new insights in cognitive development based on this approach. The core concepts of cognitive developmental robotics are "physical embodiment," and more importantly, "social interaction" the latter of which yields an information structure through interactions with other agents. Cognitive development is thought to seamlessly involve both of the above [21], [22]. In the case of early vocal development, the main issue is to identify the correspondence between the utterances of an infant and its caregiver, beyond the differences in the acoustic features. Hereafter, we refer to this as the "correspondence problem." In this paper, the following questions are addressed in relation to the constructive approaches:

1) What kinds of (social) biases and responding behaviors affect early vocal development? [social interaction]
2) How are self (unsupervised or reinforcement) or interactive (supervised) learning methods coordinated during the interaction process, and how are they related to each other?
3) What kinds of platforms (real robots or computer simulation) are used and how? [physical embodiment]

In addition, to what extent can these approaches explain real early vocal development in infants, and what issues should be addressed in the future?

The rest of this paper is organized as follows. First, the early development of infant speech perception and articulation is briefly reviewed, using data from observational studies in developmental psychology and some neuroscientific imaging studies. Next, computational modeling using real robots and/or computer simulations is examined to address how infant–caregiver interactions affect the early development of vocalization. Finally, future issues are discussed.

Manuscript received October 8, 2015; revised January 20, 2016, March 15, 2016, and April 3, 2016; accepted April 6, 2016. Date of publication April 13, 2015; date of current version June 8, 2016. This work was supported by the Grants-in-Aid for Scientific Research under Research Project 24000012.

The author is with the Graduate School of Engineering, Osaka University, Suita 565-0871, Japan (e-mail: asada@ams.eng.osaka-u.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCDS.2016.2552493

2379-8920 © 2016 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

II. Early Vocal Development: Behavioral and Neurophysiological Studies

In developmental psychology, it is claimed that infant–caregiver interaction plays an important role in an infant's vocal development, including the prelinguistic period [23]. Here, the studies related to the development of an infant's perception and articulation are briefly reviewed along with a few studies in neurophysiology.

A. Behavioral Studies on Development of Speech Perception

In general, the listening ability of an infant, with regard to adult voices, is independent of the native language of the caregiver at birth, but gradually adapts to the caregiver's native language [24]. A newborn baby can discriminate its mother's voice from others, which indicates that the prenatal auditory experience influences postnatal auditory preferences [25].

The perceptual magnet effect is a psychological phenomenon whereby a person perceives a stimulus as a prototypical one, close to one of the categories that the person has. This effect can be observed in infants who have reached around six months of age [26]. Kuhl *et al.* [27] reported that infants younger than six months can discriminate between vowels in any language, but their perception is gradually tuned to their native language, and they appear to lose the universal perceptual capability before they are six months old.

Based on the results of experiments using synthetic sounds, Kuhl [28], [29] suggested that preverbal infants can categorize speech sounds, a requirement for infants to develop the ability to perceive and produce speech. Based on an observation that three-month-old infants imitated their mothers' utterances even though these had a fundamental frequency that was different from that of the infants, Lieberman [30] suggested that three-month-old infants may be capable of vocal tract length normalization.

Based on the results of the above studies, it appears that infants might be able to identify the correspondence of vowel categories between their own utterances and those of their caregivers' despite the difference in their frequencies at an age of up to six months at the latest, and three months at the earliest, which implies that an infant's learning of vowel categories starts during the cooing period.

B. Behavioral Studies on Development of Articulation

In terms of developments that eventually lead to speech production, infants' utterances are initially (at 12 weeks) vowel-like sounds, which change to well-separated vowels (at 20 weeks) as shown in [31, Fig. 3]. They claimed that an infant's ambient language experiences (perception) influence his or her speech production by extending their natural language magnet model. During this period, changes in the shape of the oral cavity and improvements in tongue movements are observed [32], along with the descent of the epiglottis [33].

Oller [34] proposed five preverbal developmental stages for an infant's speech (phonetic control) during the first 12 months, from quasi-resonant nucleus production to variegated babbling. Nathani *et al.* [35] modified and extended Oller's model [34] to the developmental stages up until 18 months, allowing slight overlaps. These are as follows:

1) *level 1:* reflexive (around 0–2 months);
2) *level 2:* control of phonation (around 1–4 months);
3) *level 3:* expansion around (3–8 months);
4) *level 4:* basic canonical syllables around (5–10 months);
5) *level 5:* advanced forms around (9–18 months).

These processes are supposed to be supported by the development of the central nervous system controlling the rhythmic movements and the muscles used to control vocal tract movements.

C. Behavioral Studies on Development of Vocal Interaction

From the first month after birth, a mother's speech aimed at her infant is different from that of normal adult speech. That is, it is high in pitch and has many other features that are more pronounced than in normal adult-directed speech. Such speech is called "motherese," "parentese," "infant-directed speech (IDS)," or "baby talk" (hereafter, IDS). Fernald [36] found that four-month-old infants prefer to listen to IDS. Based on the results of measuring speech discrimination in infants (aged 6–8 months and 10–12 months), Liu *et al.* [37] found that the clarity of maternal speech directly affects an infant's early language learning. In addition, Werker *et al.* [38] showed that IDS contains language-specific information to establish native vowel categories.

Confirming the early observations made by Pawlby [39], Kokkinaki and Kugiumutzakis [40] reported an important characteristic of a caregiver's behavior that facilitates an infant's learning of the correspondences between the caregiver's vowels and those of the infant: parents imitate their infants much more frequently in the first six months than they do after this period.

Two kinds of infant–adult interaction experiments have been conducted [41]. One utilized conversational turn-taking, while the other addressed the random responsiveness of an adult. The results showed that an infant pronounced syllabic/vocalic sounds more frequently when the adult maintained the turn taking rather than a random pattern. As implied from the observations that an infant's vowel-like utterances prompt imitation by the caregiver [42], and that this encourages such utterances [43], parental imitation or mutual imitation might have an important role in the development of vocal imitation [44].

We may summarize the above as follows. Around the age of three months, when infants begin to learn how the vowels they produce correspond to those produced by their caregivers, infants merely produce immature cooing. Nevertheless, the caregivers respond to these vowel-like utterances, and the infants respond back with imitation. Consequently, an infant's cooing is entrained into their caregivers' vowel-like utterances. This suggests that the infant–caregiver interactions triggered by the caregivers' imitation may play a role in teaching the vowel correspondence between the infant's and caregiver's speech.

D. Neuropsychological Studies

Recently, the application of imaging technology has shown that an infant's brain exhibits a perception of language before the onset of speech production. Dehaene-Lambertz *et al.* [45] utilized functional magnetic resonance imaging to detect the brain regions responsible for normal and reversed speech in three-month-old infants who are either asleep or awake, and suggested that the left-lateralized brain regions (precursors to the adult cortical language areas) are already activated in infants before they produce speech. Imada *et al.* [46] conducted a developmental magnetoencephalography (MEG) study in which newborns and babies aged 6 and 12 months old perceived speech and nonspeech sound stimuli. The superior temporal and inferior frontal regions of the infants of all three age groups were observed to be activated, which suggests that speech perception requires a perceptual-motor link in these early periods of development. In the first instance, speech perception does not activate the speech motor areas, such that experience is needed to associate perception and action in early speech development [46].

However, it remains difficult to investigate the links between these previously mentioned early sensitivities and caregivers' interaction throughout the course of speech and language development because the current imaging technology is limited. Recently, Hirata *et al.* [47] constructed a hyperscanning system with two MEG systems in a single magnetically shielded room to examine the brain-to-brain interactions between a child and his or her mother. Further, research utilizing this technology is expected to yield new insights. Approaches based on observation appear to be difficult because ethical problems arise in controlling infant development to investigate this question. Therefore, modeling approaches are expected to contribute to the identification of the missing links.

III. Overview of Modeling Approaches

Approaches to the modeling of early vocal development can be roughly classified into four types as shown in Fig. 1. The first two types, Fig. 1(a) and (b), deal with noninteractive cases, while the second two types, Fig. 1(c) and (d), consider interactive ones. The latter are further classified into cases of interactions between homogeneous agents and heterogeneous ones such as infant–caregiver interactions. The limitations associated with types Fig. 1(a)–(c) in relation to the correspondence problem are pointed out, and then type Fig. 1(d) is introduced to solve the correspondence problem.

A. Noninteractive or Interactive Cases With Homogeneous Agents

Explanations for three types (a), (b), and (c) in Fig. 1 are given as follows:

1) *Motor Control Ability Development Through Self-Monitoring of Vocalizations:* Guenther [2] proposed a neural network model called "directions into velocities of articulators" (DIVA), which addresses an infant's acquisition of speaking skills and their subsequent motor equivalent production of speech sounds. Kanda *et al.* [3] proposed a continuous vocal imitation system based on a recurrent neural network with parametric bias that explains how infants acquire phones.
2) *Statistical Estimation of Caregiver's Vowel Categories From Their Vocalizations:* An algorithm with expectation-maximization was proposed by Vallabha *et al.* [4] to learn vowel categories from a vowel token sequence. It does not require any category information with each vowel token, the number of categories for learning, or access to the entire data set. McMurray *et al.* [5] implemented a statistical learning mechanism in a computational model with a mixture of Gaussian's architecture to determine the sufficiency of the statistical learning hypothesis and its implication to language development. They found that statistical learning alone is not sufficient for phonetic category learning, and that an additional competition mechanism is needed to successfully learn the categories in the input successfully.
3) *Self-Organization of Shared Vowels Through Imitative Interaction:* Oudeyer [11] constructed a society of artificial agents with a mechanism for forming a discrete speech code, assuming no *a priori* linguistic capacities or coordinated interactions. De Boer and Zuidema [12] investigated the evolution of a fundamental characteristic of human speech called "combinatorial phonology" using a population of simulated agents.

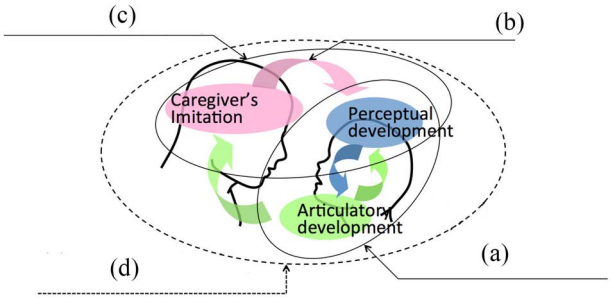

Fig. 1. Modeling approaches for vocal communication (adapted from [48]). (a) Motor control ability develops through self-monitoring of vocalizations (see [2], [3]). (b) Statistical estimation of caregiver's vowel categories from caregiver's vocalizations (see [4], [5]). (c) Self-organization of shared vowels through imitative interaction (see [11], [12]). (d) Whole dynamics of interactions (see [13]–[18]).

There are other types of computational models of speech development. One of these mainly focuses on speech perception. The EU ACORNS project[1] is one example of the systematic and extensive studies that have been conducted to acquire language and communication skills based on sensory input [49], [50]. Another type does not involve the explicit handling of the corresponding problem by assuming that various ambient auditory inputs include a caregiver's utterances (see [6]–[10]). These studies are discussed in Sections V and VI. Räsänen [51] thoroughly reviewed the computational modeling approaches that are mainly based

[1] http://lands.let.ru.nl/acorns/

on offline simulations with prerecorded data. Some of the previously mentioned approaches are included in this review.

B. Limitations of the Above Modeling Approaches

The previously mentioned modeling approaches incur several limitations. Case 1) did not consider how the learner's utterances affected the caregiver's utterances; that is, interactive communication was not examined. In case 2), the phonetic category acquired by the learner was not that of the learner but that of the caregiver; therefore, the correspondence problem was not considered, nor was its effects on the utterances of both speakers. In case 3), the multiagent society was homogeneous, that is, the agents had the same auditory systems and motor (articulation) systems. The imitation game that took place between them was based on acoustic matching which does not seem plausible in the case of infant–caregiver interactions, because they are heterogeneous agents. The sizes of the vocal tracts of adults and infants are different [52], and as a result, their sound qualities are also different from each other.

To address the correspondence problem, the overall dynamics of the vocal interaction between an infant and a caregiver should be considered, as indicated by the large broken ellipse in Fig. 1. Several studies have addressed this issue by utilizing the concept of direct mapping between the infant's and caregiver's utterances [13], the transformation of the caregiver's utterances to those of the infant [14], [53], the caregiver's affirmative biases [15], the learner's bias [16], and the caregiver's reformulation [17], [18]. These approaches are briefly reviewed in the next section.

IV. Modeling Whole Dynamics of Early Vocal Interactions

A. Loose Definition of Imitation

Gattegno [54] pointed out the problem of the loose definition of "imitation" as follows.

Here is an exercise to indicate in precise terms what we mean. Many people say "children learn to speak by imitation" and are convinced it is true. If they really want to know whether it is true, they should ask themselves, for example, what they mean by imitation. Do they mean that a baby sees what a speaker does with his throat or tongue and then reproduces these actions? Or do they mean that if this use of oneself were known to a child it would be easy for him to do what others do, although in fact he only hears people in the environment speaking and what the child must do is not hear but speak?

In the case of limb movements, for example, there might be three categorical levels of imitation.

1) *appearance level:* exactly the same trajectory should be realized;
2) *action unit level:* the same action units should be realized in the appropriate order, but the accurate trajectories of the units are not required;
3) *goal-oriented level:* the same goal should be achieved regardless of the exact means.

Imitation based on acoustic matching can be categorized into the first appearance level, but it is not a viable solution for the infant–caregiver vocal interaction. This requires an understanding of the correspondence between the utterances (vowel, consonant, or consonant plus vowel) or their combinations (words), which may correspond to the second or third level. However, this is no longer simply an issue of direct imitation, but rather of how the infant and caregiver affect each other with respect to vocal learning.

B. Approaches to Whole Dynamics of Interaction

Rochat [55] claimed that a caregiver's affirmative interpretation and imitation of an infant's immature behavior facilitates the development of the infant's social abilities. The following studies attempted to realize this by using computational models and/or real robot experiments with a variety of implementations. The key ideas of these studies are shown in Fig. 2.

Yoshikawa *et al.* [13] built a vocal robot called Burpy (see Fig. 3), which consisted of articulation and auditory parts with corresponding layers. Both layers were self-organized and connected by Hebbian learning through parrot-like teaching by a caregiver. The learner created direct mapping between the caregiver's utterances and its own articulation to avoid the corresponding problem. Based on the assumption that the learner is able to roughly estimate the mapping between the caregiver's vowel primitives and its own, Miura *et al.* [14] examined how different transformations (mappings) such as translation, rotation, scaling, and their combinations in the formant space acted to solve the corresponding problem.

A similar idea was applied in the study by Heintz *et al.* [53]. They utilized several feature vectors derived by different transformations. The relative feature ($F2$-$F1$, $F3$-$F2$: Fi denotes the ith formant frequency) was able to classify vowels, while the normal formant feature ($F1$, $F2$) was not because of the corresponding problem.

Inspired by the previous work [14], Ishihara *et al*. [15] computationally modeled an imitation mechanism as a Gaussian mixture network (GMN) in which numerous parameters were utilized to represent a caregiver's sensorimotor biases, i.e., the perceptual magnet effect [26], and an automirroring bias in the caregiver's perception, by which the caregiver perceived the infant's voice as being similar to his or her previous utterance. Both biases worked together to guide the infant's vowel categories toward the caregiver's vowel categories (Fig. 2).

The above studies assume that the caregiver almost always or always imitated the infant. However, in real situations, the rate of imitation by the caregiver is much lower. Miura *et al.* [16] addressed this issue by considering cases where the caregiver's imitation was less frequent ($<20\%$) in computer simulations using real data from the experimenter's voice recordings, and proposed a method with another automirroring bias on the infant side, which actively selects the infant's action and data using incomplete classifiers for the caregiver's imitation of the infant's utterances (the bottom-right in Fig. 2).

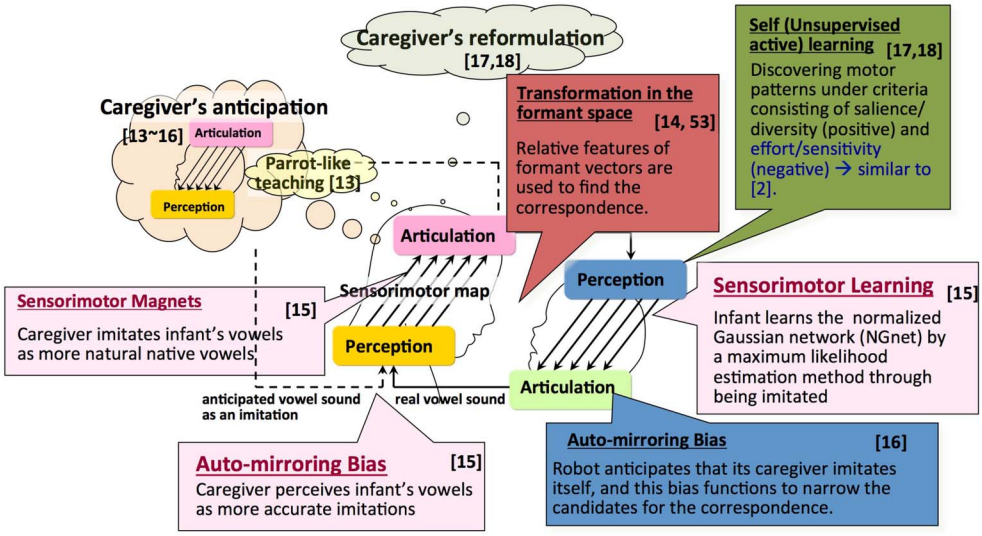

Fig. 2. Key terms for vocal interaction between an infant and a caregiver (adapted from [48]).

Following the whole dynamics concept in these studies [13]–[16], and inspired by the work of Gattegno [54], Howard and Messum [17], [18], [56] addressed the issue using Elija, a virtual infant based on a computational model. In the latest version of Elija, through active self-learning, Elija first discovered the motor patterns of sounds. Next, native speakers of English, French, and German interacted with Elija as its caregiver, and Elija memorized the caregivers' responses and reacted to the memorized patterns. This interaction was expanded to word teaching. Fig. 4 shows an example of reformulation by a caregiver in word teaching, which corresponds to the word learning by Elija, the architecture of which is shown on the right. Howard and Messum's [18] results demonstrated that human subjects naturally behaved and responded to infant-like vocalizations (Elija), and that this could take Elija from the stages of cooing/babbling to word pronunciation.

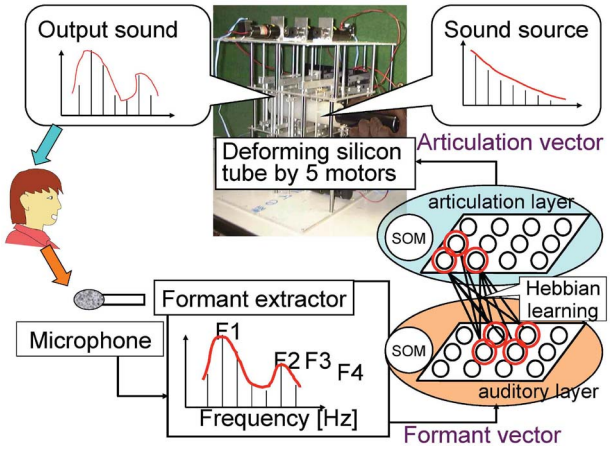

Fig. 3. Overview of the system of Burpy (adapted from [13]).

V. Relationships Between Approaches to Modeling Whole Dynamics of Interactions

Considering the developmental process of speech perception and articulation, we now discuss the following issues pertaining to the above studies of whole interaction dynamics.

A. Coordination of Self and Interactive Learning Methods

As we mentioned in Section II-B, developmental changes in an infant's vocalization ability can be observed from levels 1 (reflexive) to 5 (advanced forms). In addition to the subcomponents of communication such as perceptual biases; cognitive, anatomical, and physiological substrates; social context, mainly interaction with a caregiver, plays a key role in the vocal development [35]. Even in the case of nonhuman primates (marmoset monkeys), a caregivers' feedback facilitates the development from immature (e.g., crying) to mature vocalization, and body development alone cannot explain such vocal development [57]. However, computational modeling of these processes seems difficult because the communication subcomponents and the social context are difficult to separate.

Constructive approaches adopt several styles to model this process. One assumes separate processes for self-learning and interaction, whereas another mixes the two from the first instance or does not include self-learning (primitives are given *a priori*). The following items provide additional explanations.

1) *Separate Processes of Self-Learning and Interaction:* Elija [18] adopted this type of learning for the convenience of computation and to make it possible to analyze the roles of the behaviors of different learning schemes across multiple caregivers. Optimization criteria were applied that consisted of salience/diversity (positive) and effort/sensitivity (negative) terms. The salience/diversity terms encouraged Elija to find motor patterns, particularly novel ones. Similar criteria were used in Miura *et al.*'s study [16] of how the utterances by a learner are selected by the caregiver. The effort represents the total energy consisting of the movement

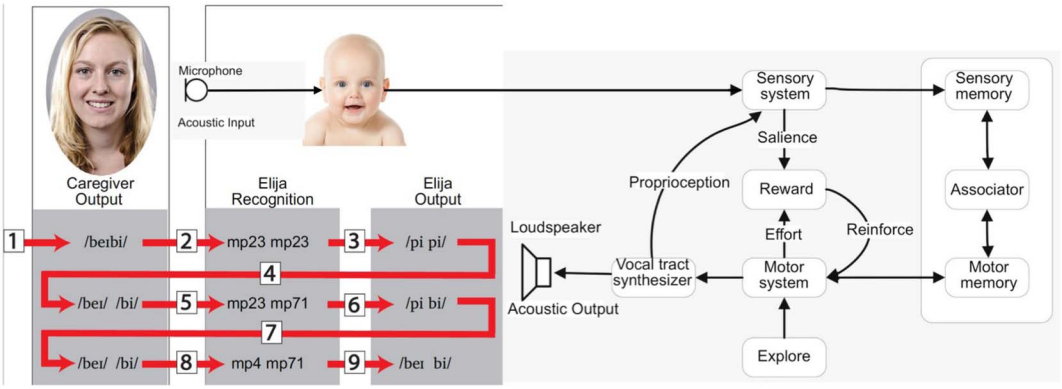

Fig. 4. Example of reformulation in word teaching (left), and Elija's architecture (right) (adapted from [17] and [18]).

of the vocal tract and the volume of the utterance. This criterion was first referred to as "toil" by Burpy, and indicated the deformation of the vocal tract [13]. The sensitivity term is also related to toil, i.e., to penalize the discovery of motor patterns with very accurate articulations. Toil almost corresponds to the term "effort" as used in the criterion of Elija's self-learning process [18]. After the self-learning, Elija fixed its articulation patterns. In other words, Elija kept its motor patterns because real-time interaction with its caregivers was possible while it pruned away poor ones. Only the initial self-learning stage optimized Elija's articulation patterns. Because of the recent progress in vocal simulation, Elija acquired a reasonable set of phones.

2) *No Self-Learning Process:* Miura *et al.* [16] focused on the selection process and finding the correspondence between the vowels with initially fixed motor patterns without any self-learning process at the beginning. Their method can be regarded as being akin to Elija's process after self-learning because the learned (Elija) or initially fixed (Miura) motor patterns did not change during the process of interaction with the caregiver. In contrast, Burpy [13] used an interaction process from the beginning without the self-learning process. Although the motor patterns were fixed at the beginning [discrete motor commands for vocal robots with five degrees of freedom (DOF)], the final motor patterns after the interaction converged to certain values in the continuous motor space, which were often shifted from the initial fixed ones. In the case of Ishihara *et al.*'s simulation [15], motor patterns were represented in the continuous space as a GMN for both the learner and the caregiver, and the learner's parameters changed during the interaction and converged to the desired values depending on the caregiver's bias parameters.

As Howard and Messum [18] mentioned, the self-learning and interaction processes could occur in parallel. Elija can have two simultaneous processes and it is possible to improve the motor patterns. However, the system could be more complicated. In Burpy [13], [15], two processes could be realized alternatively, that is, day-time interaction and night-time mental rehearsal. In the real situation of the developmental process of a human infant, these two processes could occur in parallel. However, fixed motor patterns are probably used more often during the early period, with their use gradually shifting to learned patterns. This is similar to joint attention learning [58], where the innate visual attention mechanism acts first, but after which the results of joint attention learning gradually begin to be used more often. One common issue is how to switch between the two modules, for example, based on a fixed gradual change or depending on some criterion such as the tolerance of the performance in early learning.

There are also other types of computational models based on intrinsic motivation or social rewards. The former is related to self-learning, and the latter to the affirmative bias. Moulin-Frier *et al.* [6] proposed a general exploration mechanism with intrinsic motivation, or in other words, "curiosity-driven learning," by which an agent can self-organize early vocal development through auditory interaction with a caregiver. This mechanism can explain how the learning proceeds from vocal self-exploration almost independently of ambient speech to more socially influenced vocal exploration. A spiking neural network model was proposed to control the lip and jaw muscles of an articulatory speech synthesizer and learn canonical babbling [59]. The model showed that self-learning based on self-motivated intrinsic reinforcement, and affirmative bias as social reinforcement work together for humans to acquire their canonical babbling.

Although the corresponding problem caused by the acoustic feature difference between an infant and a caregiver has not been solved, Murakami *et al.* [7] proposed a reverse order of the learning, that is, supervised learning (target selection to imitate) first, and then self (reinforcement) learning by error minimization. Because an infant (or even fetus) is exposed to various auditory inputs from its caregiver (mother), the first self (unsupervised) learning may include ambient auditory data unconsciously given by the caregiver or other adults.

Such modeling without explicitly assuming the caregivers interaction can be observed in the study by Westermann and Miranda [8] who proposed a computational model of the effects of sensory-motor mappings on the perception of vocalizations. Self-produced sound and heard (ambient) sound were analyzed together.

B. Caregiver's Affirmative Bias

One of the core ideas of cognitive developmental robotics is social interaction, which refers to the issue of the kinds of behavior by a caregiver that affect an infant's responses and how this influences the learning of speech perception and articulation. Parrot-like teaching was the caregiver's behavior toward Burpy [13], where the caregiver tended to interpret the infant robot's random cooing as her own vowels as often as possible to accelerate the learning (finding correspondences). As a result, each of the vowels acquired by the robot had a large variance because of the caregiver's tendency. To reduce the variance, subjective criteria such as smaller energy consumption and less deformation of the vocal tract were introduced, and the robot's vowels were successfully converged. The caregiver's tendency was a kind of caregiver's affirmative bias, even though objective support had not been given.

To cope with less imitative caregivers, a self-evaluation mechanism [16] with an automirroring bias on the learner's side rejected incorrect mappings (with less clear robot voices) as outliers, on the assumption that the caregiver responded more consistently to the clearer (easier to imitate) robot voices, which seemed to be a sort of affirmative bias by the caregiver (unconsciously).

Native speakers of English, French, and German interacted with Elija as its caregiver during the experiment. First, they were asked to naturally respond to Elija's utterances when they felt these were natural. Howard and Messum [18] applied a phonemic transcription analysis to the caregivers' utterances during the interactions, and it turned out that Elija's output was interpreted by the caregivers within the frameworks of their native languages. These data appeared to be objective in terms of showing a caregiver's affirmative bias, although the number of subjects was small.

Such affirmative biases of the caregivers were formalized by Ishihara *et al.* [15]. The first one arises from "sensorimotor magnets," by which a caregiver perceives and imitates infant vocalizations as if they were prototypical vowels of the caregiver's native language. The second is the automirroring bias, in which a caregiver hears the infant's vocalization as being much closer to the expected vowel because the caregiver anticipates imitation by the infant. Caregiver–infant interaction was computationally simulated, and they found that the sensorimotor magnets worked to compose small clusters, and the automirroring bias refined these clusters to make the vowels clearer. As a result, both were needed for the infant to learn the vowels of the caregiver's native language.

C. Learner's Strategies

Burpy [13] has a learning module consisting of an auditory layer and an articulation layer. These two layers collect the formant features given by the caregiver and the articulation vectors corresponding to the vowels of the caregiver's native language, which are found during the caregiver's parrot-like teaching. In each layer, self-organizing mapping is applied to find clusters, and these clusters are associated by Hebbian learning between the two layers. The Hebbian learning was modified slightly so that the size of the final cluster could be made small by introducing the previously mentioned toil parameter criterion (less deformation and less energy consumption). After learning, this association plays the role of a mirror neuron system, allowing Burpy to access its articulation vector when one of the caregiver's vowels is heard.

Ishihara *et al.* [15] represented the learner's vowel primitives as a GMN, and its parameters changed during the interactions with a caregiver, which indicated the developmental process of finding the correspondence of the vowels in the speech of the learner and caregiver.

Howard and Messum [18] reported that their caregivers' (four English, two German, and two French) responses were almost always reformulations (more than 90%) and did not contain much mimicry (less than 10%), except in the case of one English caregiver whose responses were 60% reformulation and 40% mimicry. Therefore, Elija utilized a strategy to memorize the patterns in the responses of its caregiver and responds to her with the most similar pattern. Through many cycles of such feedback, Elija is expected to statistically converge its responses and to consolidate its memory patterns to respond appropriately.

TABLE I
SUMMARY OF APPROACHES TO WHOLE DYNAMICS OF INTERACTION

	Utterance	Caregiver	Learner	Key idea to solve the correspondence problem
Yoshikawa et al., 2003 [13]	V	real	real	direct mapping (no self auditory feedback)
Miura et al. 2007 [14]	V	real	real	transformation in the formant space & lip shape imitation
Heintz et al., 2009 [53]	V	virtual	virtual	relative features of the formant
Ishihara et al. 2009 [15]	V	virtual	virtual	caregiver's sensorimotor magnet and auto-mirroring biases
Miura et al. 2012 [16]	VV	real (much less imitative)	virtual	learner's auto-mirroring bias and the caregiver's preference to easier (clearer) voices to imitate
Howard & Messum, 2011, 2014 [17,18]	V, CV, CVCV	real	virtual	separated self-learning and caregiver's reformulation

D. Interaction Methods

Table I summarizes the approaches to the whole dynamics of early vocal interaction, where V, C, CV, and CVCV stand for a vowel, a consonant, a vowel + a consonant = a syllable, and two syllables = a word, respectively.

Single-vowel mutual imitation was assumed in [13]–[15] and [53]. This could be the cooing process for infant vocalization. However, it does not seem realistic for the caregiver to always respond using a single vowel. The actual rate of the exact imitation response was less than 20% [60]. Therefore, an examination of these methods should be conducted to determine how their performances degenerate when there is such a low rate of imitation case. Miura *et al.* [16] proposed an automirroring bias on the learner side to cope with a less frequently imitative caregiver. Reformulation by the caregivers [17], [18]

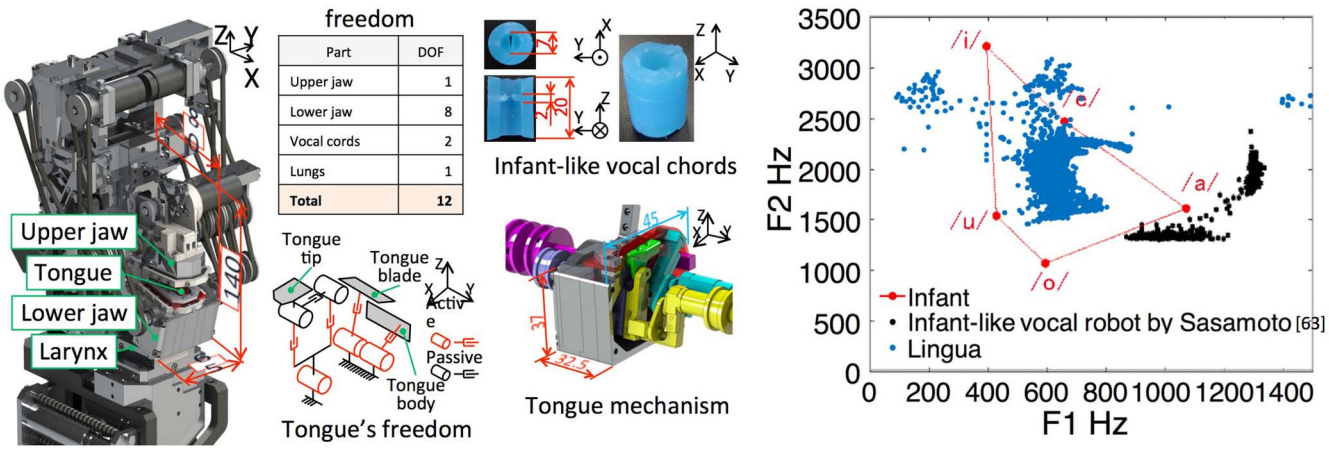

Fig. 5. Lingua: overall appearance, vocal cords and tongue, utterance range in the formant space (adapted from [62], [63]).

seems to be another realistic response to entrain the infant utterances into those of the caregivers.

On the other hand, an exaggerated articulation of IDS produced by a caregiver facilitates infant vocal learning [61]. This appears as a transformation of a caregiver's utterances to an infant's utterance region (translation and scaling) as done by Miura *et al.* [14]. However, how IDS occurs and how it affects an infant's learning are essential issues to be addressed.

Many systems have used only an auditory channel, with a few adding vision, such as Murakami *et al.* [7] and Miura *et al.* [14], who reported that the visual input facilitated vowel learning. Although this does not seem surprising, multimodal interactions should be further investigated to study how different modalities interact to promote vocalization learning of a caregiver's native language to cope with more realistic situations.

E. Research Platforms

Physical embodiment is one of the core ideas of cognitive developmental robotics [19], and in the case of vocal interaction, there are two types: 1) physical articulation systems; and 2) virtual ones. These are important parts of all the previously mentioned systems.

The "source-filter theory of speech production" [64] explains vocalization as the output of a filter function that modulates the source of sound energy in terms of the shape of the vocal tract. Based on this theory, Yoshikawa *et al.* [13] constructed Burpy. It uses an artificial larynx vibrator as a sound source, which is actually a medical tool for patients with damaged vocal chords, along with a silicone tube as a simplified vocal tract, the shape of which can be changed using five electric motors. This mechanism is similar to that of Higashimoto and Sawada [65]. They applied a conventional artificial vocal chords, but Yoshikawa *et al.* [13] used a membrane with a vibrator which oscillated at the fundamental frequency.

Miura *et al.* [14] improved Burpy as follows:

1) replacement of the sound source with an air compressor and an artificial vocal band;
2) addition of a lip at the front end of the vocal tract;
3) length reduction of the vocal tract from 170 mm (average vocal tract length of a human male) to 116 mm.

Burpy and its modification are still simple approximations of the human vocal system, because the vocal tract is a silicone tube, whereas the actual shape of a human vocal tract is much more complicated. A pioneering work on an artificial vocal system based on the anatomical knowledge of human vocalization was the series of Waseda Talkers[2] designed to mechanically reproduce the human speech in 3-D. An animatronic model of the human tongue and vocal tract, called "AnTon," was designed by Hofe and Moore [66]. They reproduced human speech gestures based on AnTon's tongue control.[3] Because their main purpose was to investigate animatronic control, the quality of the reproduced sounds was not discussed. The most recent version of Waseda Talker is WT-7R [67], which offers improved tongue performance, resulting in clearer vowels, and a sharper bandwidth for the formant peak in the spectral data.

Since the WT series models adult vocalization, Sasamoto *et al.* [68] designed a vocal robot with an infant-like articulatory system that has one DOF for the velum and jaw, two DOFs for the vocal chords, and four DOFs for the tongue. They found that regardless of the anatomical shape similarity, its vocalization performance was poor because of its having fewer DOFs for vocalization control. Therefore, a redesign of the actuation system is needed to allow the robot to vocalize with a sufficient number of DOFs as an infant. Endo *et al.* [62], [63] developed an infant-like vocal robot, Lingua, as a vocal robot platform that affords a model of real infant vocalization (Fig. 5). Lingua can produce an infant-like voice and has a high articulation capability. The shapes of its vocal chords and vocal tract are similar to those of a six-month-old infant as determined from anatomical data. Seven DOFs of tongue articulation were realized using a sophisticated design consisting of linkage mechanisms inside a miniaturized vocal tract, and this enabled

[2]Prof. Takanishi's laboratoy has been developing several vocal robots that mechanically generate sounds with artificial vocal chords.

[3]Visit https://www.youtube.com/watch?v=ZFT9B6DT6w.

the achievement of a high articulation performance (at the bottom-left in Fig. 5). The relationship between the material hardness of the vocal fold and its acoustic performance was examined, and the preliminary experiments can be seen in the right-hand part of Fig. 5, where a typical infant Japanese infant's vowels are indicated as a red pentagon, while the utterances by Lingua are indicated as small blue circles. The performance of Lingua fits the infant vowel region better than the small black circles which correspond to the utterances produced by a vocal robot [68]. Lingua needs additional improvements but will soon be used for experiments on interactions with human caregivers.

The recent progress made in articulation simulator technology in terms of the anatomical structure, function, and motor control is striking (see [69]). Elija's motor control system [18] incorporates a Maeda [70], [71] articulatory speech synthesizer. A motor pattern consists of a sequence of articulation targets, allowing the synthesizer to control ten parameters, which are interpolated, assuming that the trajectories of the articulator movements are given by a second-order critical damping equation. The synthesizer is driven by the sequences of computed time-varying parameter vectors, and the resulting acoustic output is generated through a loudspeaker. The sound quality has become very close to being human-like, which enables Elija to interact with human subjects in realtime.

It has been suggested that articulation simulators are not good at generating real-time responses to human subjects. Elija has partially solved this issue by conducting the self-learning process offline and selecting one fixed motor pattern during real-time interactions. Real vocal robots are expected to exhibit natural acoustic properties based on the real, fluid dynamics of the airflow. Moreover, natural acoustic transitions between syllables or words are possible because of the physical movements with different soft materials. It seems difficult for articulation simulators to generate such acoustic transitions because they only produce memorized motor patterns, regardless of contextual the information. These issues might be resolved in the future, but the most important issue to be addressed is how these aspects affect the caregivers' responses, which also affect the learning performance.

VI. Conclusion

In the previous section, we discussed the relationship between the approaches to whole dynamics of the interaction between an infant and a caregiver. One of the issue is the lack of a neuroscientific perspective because of the difficulties of infant brain imaging and explicitly handling the interaction issue itself. Therefore, neuroscientific approaches deal with this implicitly by regarding a caregiver's utterances as an ambient auditory input [8]. Kröger *et al.* [9] simulated speech acquisition, production, and perception based on the cortical, subcortical, and peripheral sensorimotor mapping structure. Their older version of their model [10] contained more details about the neural computations for the functions in the corresponding brain regions. Guenther *et al.* [72] also extended the DIVA model [2] to a neural model supported by imaging studies, but with less emphasis on ambient auditory input. These approaches mainly focused on the learner's internal mechanism based on neuroscientific knowledge. Unfortunately, these models seem to be based on adult brains because imaging studies of immature infant brain development are still difficult to conduct. Nevertheless, these models and approaches are important as candidates for the internal mechanisms of infant vocal learning, and should be combined or integrated with social (interaction) learning methods to reveal how infants learn to vocalize their caregiver's native language.

One of the most serious issues is the extent to which we should follow the neuroanatomical and/or neurophysiological findings. For example, Jürgens [73] provided a review on the neural pathways underlying vocal control, and it seems almost impossible to realize vocal robots that faithfully reflect these findings. Thus, the most essential issue for constructive approaches such as for cognitive developmental robotics is to find a basic principle that can be shared by natural and artificial systems, which should contribute to the acquisition of new insights into early vocal development, and more generally human cognitive development.

Issues to be addressed in the future can be summarized as follows.

1) Integration of neuroscientific approaches focusing on neural mechanism inside the learner and interactive ones focusing on social learning issues. The extent to which the system is neuroanatomically and/or neurophysiologically implemented could be related to which social learning aspect is considered.
2) *More Realistic Interactions:* The relationships between multimodal sensations, not only auditory, but also vision and touch should be analyzed. The developmental change in these relations (cooperative, interfering, or independent) is an interesting topic.
3) *More Experiments With Real Humans:* Systematic experiments with real human subjects should be performed to verify and improve the models. Further, IDS issues should be considered.

Although interdisciplinary approaches are the minimal requirement, it is more important to find a principle shared by different disciplines and its contribution to the gaining of new insights.

Acknowledgment

The author would like to thank Prof. K. Hosoda, Dr. Y. Yoshikawa, and Dr. H. Ishihara with Osaka University, and Dr. K. Miura and Dr. Y. Sasamoto with Fujitsu Company for their support in attaining the achievements described in this paper. The author would also like to thank the productive comments of the three reviewers that helped improve this paper.

References

[1] T. W. Deacon, *The Symbolic Species: The Co-Evolution of Language and the Brain*. New York, NY, USA: W. W. Norton & Company, 1998.
[2] F. H. Guenther, "A neural network model of speech acquisition and motor equivalent speech production running title: Speech acquisition and motor equivalence," *Biol. Cybern.*, vol. 72, no. 1, pp. 43–53, Nov. 1994.

[3] H. Kanda, T. Ogata, T. Takahashi, K. Komatani, and H. G. Okuno, "Continuous vocal imitation with self-organized vowel spaces in recurrent neural network," in *Proc. IEEE Int. Conf. Robot. Autom.*, Kobe, Japan, May 2009, pp. 4438–4443.

[4] G. K. Vallabha, J. L. McClelland, F. Pons, J. F. Werker, and S. Amano, "Unsupervised learning of vowel categories from infant-directed speech," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 33, pp. 13273–13278, 2007.

[5] B. McMurray, R. N. Aslin, and J. C. Toscano, "Statistical learning of phonetic categories: Insights from a computational approach," *Develop. Sci.*, vol. 12, no. 3, pp. 369–378, 2009.

[6] C. Moulin-Frier, S. M. Nguyen, and P.-Y. Oudeyer, "Self-organization of early vocal development in infants and machines: The role of intrinsic motivation," *Front. Psychol. (Cogn. Sci.)*, vol. 4, Jan. 2014, Art. no. 1006.

[7] M. Murakami, B. Kröger, P. Birkholz, and J. Triesch, "Seeing [u] aids vocal learning: Babbling and imitation of vowels using a 3D vocal tract model, reinforcement learning, and reservoir computing," in *Proc. 5th Int. Conf. Develop. Learn. Epigenet. Robot.*, Providence, RI, USA, 2015, pp. 208–213.

[8] G. Westermann and E. R. Miranda, "A new model of sensorimotor coupling in the development of speech," *Brain Lang.*, vol. 89, no. 2, pp. 393–400, 2004.

[9] B. J. Kröger, J. Kannampuzha, and E. Kaufmann, "Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception," *EPJ Nonlin. Biomed. Phys.*, vol. 2, no. 2, pp. 1–28, Dec. 2014.

[10] B. J. Kroger, J. Kannampuzha, and C. Neuschaefer-Rube, "Towards a neurocomputational model of speech production and perception," *Speech Commun.*, vol. 51, no. 9, pp. 793–809, 2009.

[11] P.-Y. Oudeyer, "The self-organization of speech sounds," *J. Theor. Biol.*, vol. 233, no. 3, pp. 435–449, 2005.

[12] B. de Boer and W. Zuidema, "Multi-agent simulations of the evolution of combinatorial phonology," *Adapt. Behav.*, vol. 18, no. 2, pp. 141–154, 2010.

[13] Y. Yoshikawa, J. Koga, M. Asada, and K. Hosoda, "A constructivist approach to infants' vowel acquisition through mother-infant interaction," *Connect. Sci.*, vol. 15, no. 4, pp. 245–258, 2003.

[14] K. Miura, Y. Yoshikawa, and M. Asada, "Unconscious anchoring in maternal imitation that helps find the correspondence of a caregiver's vowel categories," *Adv. Robot.*, vol. 21, no. 13, pp. 1583–1600, 2007.

[15] H. Ishihara, Y. Yoshikawa, K. Miura, and M. Asada, "How caregiver's anticipation shapes infant's vowel through mutual imitation," *IEEE Trans. Auton. Mental Develop.*, vol. 1, no. 4, pp. 217–225, Dec. 2009.

[16] K. Miura, Y. Yoshikawa, and M. Asada, "Vowel acquisition based on an auto-mirroring bias with a less imitative caregiver," *Adv. Robot.*, vol. 26, nos. 1–2, pp. 23–44, 2012.

[17] I. S. Howard and P. Messum, "Modeling the development of pronunciation in infant speech acquisition," *Motor Control*, vol. 15, no. 1, pp. 85–117, 2011.

[18] I. S. Howard and P. Messum, "Learning to pronounce first words in three languages: An investigation of caregiver and infant behavior using a computational model of an infant," *PloS One*, vol. 9, no. 10, 2014, Art. no. e11034.

[19] M. Asada *et al.*, "Cognitive developmental robotics: A survey," *IEEE Trans. Auton. Mental Develop.*, vol. 1, no. 1, pp. 12–34, May 2009.

[20] A. Cangelosi and M. Schlesinger, *Developmental Robotics—From Babies to Robots*. Cambridge, MA, USA: MIT Press, 2015.

[21] M. Asada, "Can cognitive developmental robotics cause a paradigm shift?" in *Neuromorphic and Brain-Based Robots*, J. L. Krichmar and H. Wagatsuma, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2011, pp. 251–273.

[22] M. Asada, "Towards artificial empathy," *Int. J. Soc. Robot.*, vol. 7, no. 1, pp. 19–33, 2015.

[23] M. H. Goldstein and J. A. Schwade, "Social feedback to infants' babbling facilitates rapid phonological learning," *Psychol. Sci.*, vol. 19, no. 5, pp. 515–523, 2008.

[24] J. F. Werker and R. C. Tees, "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," *Infant Behav. Develop.*, vol. 7, no. 1, pp. 49–63, 1984.

[25] A. J. DeCasper and M. J. Spence, "Prenatal maternal speech influences newborns' perception of speech sounds," *Infant Behav. Develop.*, vol. 9, no. 2, pp. 133–150, 1986.

[26] P. K. Kuhl, "Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not," *Percept. Psychophys.*, vol. 50, no. 2, pp. 93–107, 1991.

[27] P. K. Kuhl, K. A. Williams, F. Lacerda, K. N. Stevens, and B. Lindblom, "Linguistic experience alters phonetic perception in infants by 6 months of age," *Science*, vol. 255, no. 5044, pp. 606–608, 1992.

[28] P. K. Kuhl, "Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories," *J. Acoust. Soc. America*, vol. 66, no. 6, pp. 1668–1679, 1979.

[29] P. K. Kuhl, "Perception of auditory equivalence classes for speech in early infancy," *Infant Behav. Develop.*, vol. 6, nos. 2–3, pp. 263–285, 1983.

[30] P. Lieberman, "On the development of vowel production in young children," in *Child Phonology*, vol. 1, G. H. Yeni-Komshian, J. F. Kavanage, and C. A. Ferguson, Eds. New York, NY, USA: Academic Press, 1980, pp. 113–142.

[31] P. K. Kuhl and A. N. Meltzoff, "Infant vocalizations in response to speech: Vocal imitation and developmental change," *J. Acoust. Soc. America*, vol. 100, pp. 2415–2438, Oct. 1996.

[32] R. D. Kent, "Sensorimotor aspects of speech development," in *Development of Perception*, vol. 1, R. N. Aslin, J. R. Alberts, and M. R. Petersen, Eds. New York, NY, USA: Academic Press, 1981, pp. 161–189.

[33] C. T. Sasaki, P. A. Levine, J. T. Laitman, and E. S. Crelin, Jr., "Postnatal descent of the epiglottis in man. A preliminary report," *Archives Otolaryngol.*, vol. 103, no. 3, pp. 169–171, 1977.

[34] D. K. Oller, *The Emergence of the Sound of the Speech in Infancy*. New York, NY, USA: Academic Press, 1980, pp. 93–112.

[35] S. Nathani, D. J. Ertmer, and R. E. Stark, "Assessing vocal development in infants and toddlers," *Clin. Linguist. Phonet.*, vol. 20, no. 5, pp. 351–369, 2006.

[36] A. Fernald, "Four-month-old infants prefer to listen to motherese," *Infant Behav. Develop.*, vol. 8, no. 2, pp. 181–195, 1985.

[37] H.-M. Liu, P. K. Kuhl, and F.-M. Tsao, "An association between mothers' speech clarity and infants' speech discrimination skills," *Develop. Sci.*, vol. 6, no. 3, pp. F1–F10, 2003.

[38] J. F. Werker *et al.*, "Infant-directed speech supports phonetic category learning in English and Japanese," *Cognition*, vol. 103, no. 1, pp. 147–162, 2007.

[39] S. J. Pawlby, "Imitative interaction," in *Studies in Mother-Infant Interaction*, H. R. Schaffer Ed. New York, NY, USA: Academic Press, 1977, pp. 203–224.

[40] T. Kokkinaki and G. Kugiumutzakis, "Basic aspects of vocal imitation in infant-parent interaction during the first 6 months," *J. Reprod. Infant Psychol.*, vol. 18, no. 3, pp. 173–187, 2000.

[41] K. Bloom, A. Russell, and K. Wassenberg, "Turn taking affects the quality of infant vocalizations," *J. Child Lang.*, vol. 14, no. 2, pp. 211–217, 1987.

[42] N. Masataka and K. Bloom, "Acoustic properties that determine adult's preference for 3-month-old infant vocalizations," *Infant Behav. Develop.*, vol. 17, no. 4, pp. 461–464, 1994.

[43] M. Pélaez-Nogueras, J. L. Gewirtz, and M. M. Markham, "Infant vocalizations are conditioned both by maternal imitation and motherese speech," *Infant Behav. Develop.*, vol. 19, no. 1, p. 670, 1996.

[44] S. S. Jones, "Imitation and empathy in infancy," *Cogn. Brain Behav.*, vol. 13, no. 4, pp. 391–413, 2009.

[45] G. Dehaene-Lambertz, S. Dehaene, and L. Hertz-Pannier, "Functional neuroimaging of speech perception in infants," *Science*, vol. 298, no. 5600, pp. 2013–2015, 2002.

[46] T. Imada *et al.*, "Infant speech perception activates Broca's area: A developmental magnetoencephalography study," *Neuroreport*, vol. 17, no. 10, pp. 957–962, 2006.

[47] M. Hirata *et al.*, "Hyperscanning MEG for understanding mother-child cerebral interactions," *Front. Human Neurosci.*, vol. 8, no. 118, 2014, doi: 10.3389/fnhum.2014.00118.

[48] H. Ishihara. (2013). *Caregiver's Auto-Mirroring and Infant's Articulatory Development Enable Vowel Sharing*. [Online]. Available: http://www.gcoe-cnr.osaka-u.ac.jp/media/handouts/bu04_ishihara.pdf

[49] L. Boves, L. Ten Bosch, and R. Moore, "ACORNS—Towards computational modeling of communication and recognition skills," in *Proc. 6th IEEE Int. Conf. Cogn. Informat.*, South Lake Tahoe, CA, USA, 2007, pp. 349–356.

[50] L. Ten Bosch, H. V. Hamme, L. Boves, and R. K. Moore, "A computational model of language acquisition: The emergence of words," *Fundamenta Informaticae*, vol. 90, no. 3, pp. 229–249, 2009.

[51] O. Räsänen, "Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions," *Speech Commun.*, vol. 54, no. 9, pp. 975–997, 2012.

[52] H. K. Vorperian and R. D. Kent, "Vowel acoustic space development in children: A synthesis of acoustic and anatomic data," *J. Speech Lang. Hear. Res.*, vol. 50, no. 6, pp. 1510–1545, 2007.
[53] I. Heintz, M. Beckman, E. Fosler-Lussier, and L. Menard, "Evaluating parameters for mapping adult vowels to imitative babbling," in *Proc. INTERSPEECH*, Brighton, U.K., 2009, pp. 688–691.
[54] C. Gattegno, *The Universe of Babies: In the Beginning There Were No Words*. New York, NY, USA: Edu. Solut. Inc., 1973.
[55] P. Rochat, *The Infant's World*. Cambridge, MA, USA: Harvard Univ. Press, 2004, ch. 4.
[56] I. S. Howard and P. Messum, "A computational model of infant speech development," in *Proc. 12th Int. Conf. Speech Comput. (SPECOM)*, Moscow, Russia, 2007, pp. 756–765.
[57] D. Y. Takahashi *et al.*, "The developmental dynamics of marmoset monkey vocal production," *Science*, vol. 349, no. 6249, pp. 734–738, 2015.
[58] Y. Nagai, K. Hosoda, A. Morita, and M. Asada, "A constructive model for the development of joint attention," *Connect. Sci.*, vol. 15, no. 4, pp. 211–229, 2003.
[59] A. S. Warlaumont, "Salience-based reinforcement of a spiking neural network leads to increased syllable production," in *Proc. IEEE 3rd Joint Int. Conf. Develop. Learn. Epigenet. Robot. (ICDL-EpiRob)*, Osaka, Japan, 2013, pp. 1–7, doi: 10.1109/DevLrn.2013.6652547.
[60] J. Gros-Louis, M. J. West, M. H. Goldstein, and A. P. King, "Mothers provide differential feedback to infants' prelinguistic sounds," *Int. J. Behav. Develop.*, vol. 30, no. 6, pp. 509–516, 2006.
[61] P. K. Kuhl *et al.*, "Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e)," *Philos. Trans. Royal Soc. B*, vol. 363, no. 1493, pp. 979–1000, 2008.
[62] N. Endo *et al.*, "Design of an articulation mechanism for an infant-like vocal robot 'Lingua,"' in *Proc. 3rd Conf. Biomimetic Biohybrid Syst. (Living Mach.)*, Milan, Italy, 2014, pp. 389–391.
[63] N. Endo, T. Kojima, H. Ishihara, T. Horii, and M. Asada, "Design and preliminary evaluation of the vocal cords and articulator of an infant-like vocal robot 'Lingua,"' in *Proc. IEEE RAS Int. Conf. Humanoid Robots*, Madrid, Spain, 2014, pp. 1063–1068.
[64] P. Rubin and E. Vatikiotis-Bateson, "Measuring and modeling speech production" in *Animal Acoustic Communication*, S. L. Hopp, M. J. Owren, and C. S. Evans, Eds. New York, NY, USA: Springer-Verlag, 1998, pp. 251–290.
[65] T. Higashimoto and H. Sawada, "Speech production by a mechanical model: Construction of a vocal tract and its control by neural network," in *Proc. IEEE Int. Conf. Robot. Autom.*, Washington, DC, USA, 2002, pp. 3858–3863.
[66] R. Hofe and R. K. Moore, "Towards an investigation of speech energetics using 'AnTon': An animatronic model of a human tongue and vocal tract," *Connect. Sci.*, vol. 20, no. 4, pp. 319–336, Dec. 2008.
[67] K. Fukui *et al.*, "Three dimensional tongue with liquid sealing mechanism for improving resonance on an anthropomorphic talking robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, St. Louis, MO, USA, 2009, pp. 5456–5462.
[68] Y. Sasamoto, N. Nishijima, and M. Asada, "Towards understanding the origin of infant directed speech: A vocal robot with infant-like articulation," in *Proc. IEEE Int. Conf. Develop. Learn. Epigenet. Robot. (ICDL-EpiRob)*, Osaka, Japan, 2013, pp. 1–2.
[69] H. Rasilo, O. Räsänen, and U. K. Laine, "Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion," *Speech Commun.*, vol. 55, no. 9, pp. 909–931, 2013.
[70] S. Maeda, "An articulatory model of the tongue based on a statistical analysis," *J. Acoust. Soc. America*, vol. 65, no. S1, p. S22, 1979.
[71] S. Maeda, *Compensatory Articulation During Speech: Evidence From the Analysis and Synthesis of Vocal-Tract Shapes Using an Articulatory Model*. Boston, MA, USA: Kluwer Academic, 1990, pp. 131–149.
[72] F. H. Guenther, S. S. Ghosh, and J. A. Tourville, "Neural modeling and imaging of the cortical interactions underlying syllable production," *Brain Lang.*, vol. 96, no. 3, pp. 280–301, 2006.
[73] U. Jürgens, "Neural pathways underlying vocal control," *Neurosci. Biobehav. Rev.*, vol. 26, pp. 235–258, 2002.

Minoru Asada (F'05) received the B.E., M.E., and Ph.D. degrees in control engineering from Osaka University, Suita, Japan, in 1977, 1979, and 1982, respectively.

He became a Full Professor of Mechanical Engineering for Computer-Controlled Machinery with Osaka University, in 1995. Since 1997, he has been a Professor with the Department of Adaptive Machine Systems, Osaka University. Since 2013, he has been the Director of the Division of Cognitive Neuroscience Robotics, Institute for Academic Initiatives, Osaka University. He was the Research Director of the Japan Science and Technology Agency Exploratory Research for Advanced Technology ASADA Synergistic Intelligence Project in 2005 and 2012. In 2012, the Japan Society for Promotion of Science named him to serve as the Research Leader for the Specially Promoted Research Project (Tokusui) on Constructive Developmental Science Based on Understanding the Process From Neuro-Dynamics to Social Interaction.

Prof. Asada was a recipient of the 1992 Best Paper Award of IEEE/RSJ International Conference on Intelligent Robots and Systems, one of ten finalists for the 1995 IEEE Robotics and Automation Society Best Conference Paper Award in 1995, the National Institute of Science and Technology Policy, Japan Award from the National Institute of Science and Technology Policy in 2006, the Okawa Publications Prize (The Okawa Foundation) in 2007, the Good Designs Award for vivid oral conversation through acquiring language (Japan Industrial Design Promotion Organization) in 2008, and the Best Paper Award of the Robotics Society of Japan in 2009 and 2016. His team called JoiTech in the humanoid adult size league in 2013 got the championship and the best humanoid award in RoboCup 2013, Eindhoven, The Netherlands. In 1997, his team was the inaugural champion (shared with the University of Southern California), in the middle-sized league of the first RoboCup competition held in conjunction with The International Joint Conference on Artificial Intelligence'97, Nagoya, Japan. In 2001, he received a Commendation by the Minister of Education, Culture, Sports, Science and Technology, Japan Government as Persons of Distinguished Services to Enlighten People on Science and Technology. He served as the General Chair for the IEEE/RSJ 1996 International Conference on Intelligent Robots and Systems, the 2005 International Conference on Development and Learning, the tenth 2008 International Conference on the Simulation of Adaptive Behavior, and the IEEE-Robotics and Automation Society 2012 International Conference on Humanoid Robots. He has been the Founding Vice President of the RoboCup Federation since 1998 and served as the President in 2002 and 2008. He was elected as a fellow of the IEEE for Contributions to Robot Learning and Applications in 2005.